# Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback

Year: 2023 | Citations: 686 | Authors: Stephen Casper, Xander Davies, Claudia Shi, T. Gilbert, J'er'emy Scheurer

---

## Abstract

Reinforcement learning from human feedback (RLHF) is a technique for training AI systems to align with human goals. RLHF has emerged as the central method used to finetune state-of-the-art large language models (LLMs). Despite this popularity, there has been relatively little public work systematizing its flaws. In this paper, we (1) survey open problems and fundamental limitations of RLHF and related methods; (2) overview techniques to understand, improve, and complement RLHF in practice; and (3) propose auditing and disclosure standards to improve societal oversight of RLHF systems. Our work emphasizes the limitations of RLHF and highlights the importance of a multi-faceted approach to the development of safer AI systems.